NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Reliability of dissimilarity measures for multi-voxel pattern analysis

Alexander Walther [a,b,*], Hamed Nili [c], Naveed Ejaz [b], Arjen Alink [a], Nikolaus Kriegeskorte [a], Jörn Diedrichsen [b]

[a] MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, CB2 7EF, Cambridge, United Kingdom
[b] Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, United Kingdom
[c] Department of Experimental Psychology, University of Oxford, South Parks Road, OX1 3UD, Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 10 February 2015
Accepted 7 December 2015
Available online xxxx

Keywords:
fMRI
Multi-voxel pattern analysis
Representational similarity analysis
Crossvalidation
Linear discriminant
Noise normalization
Classification
Decoding
Machine learning

ABSTRACT

Representational similarity analysis of activation patterns has become an increasingly important tool for studying brain representations. The dissimilarity between two patterns is commonly quantified by the correlation distance or the accuracy of a linear classifier. However, there are many different ways to measure pattern dissimilarity and little is known about their relative reliability. Here, we compare the reliability of three classes of dissimilarity measure: classification accuracy, Euclidean/Mahalanobis distance, and Pearson correlation distance. Using simulations and four real functional magnetic resonance imaging (fMRI) datasets, we demonstrate that continuous dissimilarity measures are substantially more reliable than the classification accuracy. The difference in reliability can be explained by two characteristics of classifiers: discretization and susceptibility of the discriminant function to shifts of the pattern ensemble between runs. Reliability can be further improved through multivariate noise normalization for all measures. Finally, unlike conventional distance measures, crossvalidated distances provide unbiased estimates of pattern dissimilarity on a ratio scale, thus providing an interpretable zero point. Overall, our results indicate that the crossvalidated Mahalanobis distance is preferable to both the classification accuracy and the correlation distance for characterizing representational geometries.

Introduction

It has become increasingly popular to analyze functional magnetic resonance imaging (fMRI) data using multi-voxel pattern analysis (MVPA). In MVPA, activation patterns are analyzed using either classification (Cox and Savoy, 2003; Haxby et al., 2001) or representational similarity analysis (RSA, Kriegeskorte et al., 2008). Both approaches quantitatively measure the dissimilarity of fMRI response patterns for pairs of conditions. All possible pairwise dissimilarity values of an experiment can be assembled in a pairwise decoding accuracy matrix or representational dissimilarity matrix (RDM).

One important decision in RSA is the choice of dissimilarity measure. Popular dissimilarity measures are the percentage of correct pairwise classifications (accuracy) and continuous distance measures, such as the Pearson correlation distance, the Euclidean distance, and the Mahalanobis distance. In this paper we provide a careful evaluation of the reliability of these dissimilarity measures, i.e. how reliable a measure is over replications of the experiment.

In evaluating reliability, it is important to consider the inferential aim of the analysis. One hypothesis that a researcher may want to test is that the patterns associated with conditions A and B are more similar than those associated with conditions C and D. This hypothesis concerns only the ranks of the dissimilarities. A more specific hypothesis would be that the dissimilarity between the patterns for conditions A and B is twice as large as the dissimilarity between the patterns for C and D. Here it is necessary that the dissimilarity measure have a meaningful zero point, with zero indicating that the two patterns are not different. However, distances, by definition, are non-negative and always larger than zero if estimated from noisy data. Thus, even if the true patterns are not different, the estimated distance will be larger than zero. The noise creates a positive bias, which will rise with the noise level. As we will show in the results, the bias can be removed by crossvalidation (Allefeld and Haynes, 2014; Nili et al., 2014; Kriegeskorte et al., 2007). Crossvalidated distance estimator are unbiased, i.e. their expected value equals the true distance and is zero if the two patterns are not different (see Crossvalidation section). As a consequence, crossvalidated distance estimators enable us to interpret ratios between distances.

In this paper, we compare the reliability of the Euclidean distance, the Mahalanobis distance, and the correlation distance and study the influence of univariate and multivariate noise normalization on RDM reliability. We also consider crossvalidated versions of the Mahalanobis distance (including the linear-discriminant t value; Nili et al., 2014; Kriegeskorte et al., 2007). Finally, we compare continuous distance measures to classification accuracies from linear discriminant analysis

* Corresponding author at: MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom. Tel.: +44 1223 355294.
E-mail addresses: alexander.walther@mrc-cbu.cam.ac.uk (A. Walther), hamed.nili@psy.ox.ac.uk (H. Nili), n.ejaz@ucl.ac.uk (N. Ejaz), arjen.alink@mrc-cbu.cam.ac.uk (A. Alink), nikolaus.kriegeskorte@mrc-cbu.cam.ac.uk (N. Kriegeskorte), j.diedrichsen@ucl.ac.uk (J. Diedrichsen).

http://dx.doi.org/10.1016/j.neuroimage.2015.12.012
1053-8119/

(LDA) and support vector machines (SVM). Overall, our results strongly suggest the use of continuous crossvalidated distance estimators with multivariate noise normalization to measure brain representational dissimilarities.

Materials and methods

The Euclidean distance

In RSA, we want to calculate the distance between the activation patterns b_k and b_j, corresponding to two of k = 1,…,K conditions. An activation pattern usually consist of the regression coefficients from a general linear model (GLM), which represent the response of the voxels p = 1,…,P to condition k. The Euclidean distance between two patterns in a P-dimensional voxel space, with the activity of each voxel forming a separate dimension, is defined analogously to the familiar distance in two dimensions. The squared Euclidean distance d^2 between the two row vectors $\mathbf{b}_k$ and $\mathbf{b}_j$ is:

$$d^2_{Euclidean}(\mathbf{b}_k,\mathbf{b}_j) = \|\mathbf{b}_j-\mathbf{b}_k\|^2 = (\mathbf{b}_j-\mathbf{b}_k)(\mathbf{b}_j-\mathbf{b}_k)^T = \mathbf{cBB}^T\mathbf{c}^T \quad (1)$$

where the last term represents a compact form obtained by assembling the activation patterns into a K × P Matrix **B** and applying a 1 × K contrast vector **c**, which contains zeros except for $c_j = 1$ and $c_k = -1$.

To visualize the pattern distances, imagine each pattern as a vector extending from the origin to point $\mathbf{b}_k$, where the origin of the pattern space is usually determined by the implicit baseline estimate of the GLM. The Euclidean distance between the endpoints of two vectors is independent of the origin (Figs. 1A,B). This might be advantageous if the baseline was not reliably estimated or if it cannot be meaningfully defined.

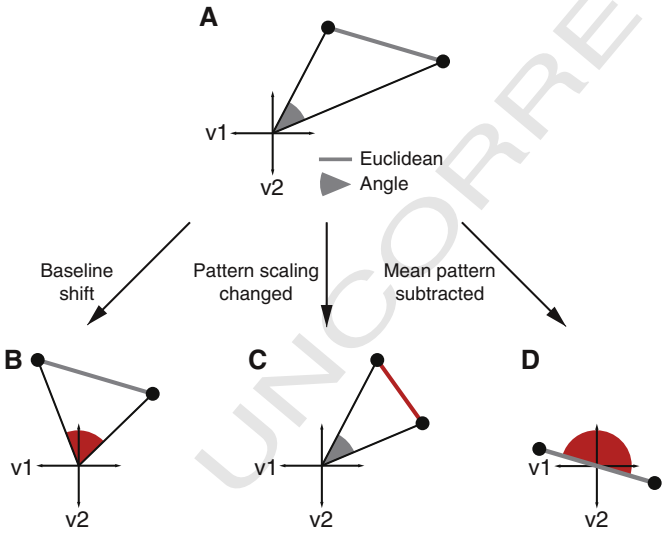

Fig. 1. Euclidean and angle-based distance in MVPA. (A) An fMRI pattern space laid out by two voxels (v1 and v2; note that the typical pattern space will often have >50 dimensions). Two pattern vectors extend from the origin. The Euclidean distance is the distance between the patterns. The cosine distance (as well as the Pearson correlation distance) measures pattern dissimilarity as a function of the angle enclosed by the vectors. (B) Shifts of the origin (i.e. the fMRI baseline) of the pattern space influence the angle (red) between the two vectors and hence the correlation distance, but not the Euclidean distance (gray). (C) Changes in the length of the two vectors (multiplicative scaling) influence the Euclidean distance (red) between the two vectors, but not the angle (gray). (D) The mean pattern of the two conditions has been subtracted (cocktail blank removal). The two vectors now extend in opposite directions from the origin, causing the cosine of the angle (red) and the correlation to become −1.

The Pearson correlation distance

Another measure of the similarity of $\mathbf{b}_k$ and $\mathbf{b}_j$ is their Pearson correlation r. The correlation is related to a slightly simpler measure, which can be more easily understood graphically: the cosine of the angle between the vectors (Fig. 1A). The cosine can be obtained by normalizing $\mathbf{b}_k$ and $\mathbf{b}_j$ by their respective L2-norms and subsequently calculating their inner product. We can then obtain a distance measure (known as cosine distance) by taking the complement:

$$d_{Cosine}(\mathbf{b}_k,\mathbf{b}_j) = 1-\frac{\langle \mathbf{b}_k,\mathbf{b}_j\rangle}{\|\mathbf{b}_k\|\|\mathbf{b}_j\|} = 1-\cos(\angle \mathbf{b}_k,\mathbf{b}_j). \quad (2)$$

The inner product detects congruent trends between $\mathbf{b}_k$ and $\mathbf{b}_j$ (i.e. when $b_{p,k}$ tends to be high, $b_{p,j}$ tends to be high as well, and vice versa). The normalization makes the cosine distance, unlike the Euclidean distance, invariant to changes in scaling (or length) of b (Fig. 1C).

The correlation distance is equivalent to the cosine distance after subtracting the mean value from each voxel pattern. If $\bar{b}$ is the voxel mean and 1 is a 1 × P row vector of ones, the correlation distance is defined as:

$$\tilde{\mathbf{b}}_k = \mathbf{b}_k-\bar{b}_k 1 \qquad \tilde{\mathbf{b}}_j = \mathbf{b}_j-\bar{b}_j 1\, d_{Correlation}(\mathbf{b}_k,\mathbf{b}_j) = 1-\frac{\langle \tilde{\mathbf{b}}_k,\tilde{\mathbf{b}}_j\rangle}{\|\tilde{\mathbf{b}}_k\|\|\tilde{\mathbf{b}}_j\|} = 1-\cos\left(\angle \tilde{\mathbf{b}}_k,\tilde{\mathbf{b}}_j\right) \quad (3)$$

The cosine and correlation distance are zero if two normalized patterns are identical. In the cosine similarity, only vector length is divisively normalized. In the correlation distance, the mean is first subtracted before divisive length normalization, making it invariant to both changes in the mean and variance of b_k across voxels. Importantly, both the cosine and correlation distance depend on the implicit baseline estimate of the GLM (Fig. 1B). Therefore, shifts in the origin will affect the overall distance structure.

The effect of mean pattern subtraction (cocktail-blank removal)

Before submitting the patterns to MVPA, it is common practice to subtract the mean pattern, i.e. the mean across conditions for each voxel, from each response pattern (Misaki et al., 2010; Op de Beeck, 2010; Pietrini et al., 2004; Williams et al., 2008, 2007). This normalization step is sometimes called "cocktail-blank removal". Removal of the mean pattern has a very different effect from removing the *mean value* (i.e. the mean of each condition, averaged across voxels, Eq. (3)). Mean pattern subtraction effectively moves the origin of the pattern space to lie in the mean pattern of all conditions (Fig. 1D). The reasoning behind this normalization step is that the response patterns may share a common component, which will increase all correlations and hence decrease the correlation distance. Mean pattern subtraction removes the influence of this common response pattern. However, the change in origin will cause unrelated patterns to be negatively correlated (Garrido et al., 2013; Diedrichsen et al., 2011). In the extreme case of only two conditions, the angle between them will always be 180 degrees and the cosine of the angle (and also the correlation) will be −1 (Fig. 1D). This can change the representational structure substantially, even when only considering the ranks of the distances. Unlike the correlation distance, the Euclidean distance is unaffected by mean pattern subtraction, as it does not depend on the origin of the coordinate system (Fig 1D).

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

Univariate and multivariate noise normalization

An important step in multivariate fMRI analysis is to take into account that signal from individual voxels is corrupted by different levels of noise. That is, some voxels may show higher signal variations than other voxels. Furthermore, noise is also spatially correlated across neighboring voxels (Friston et al., 1994; Zarahn et al., 1997).

An estimate of the structure of the noise can be obtained from the residuals of the first-level GLM. After the GLM estimation, the model residuals **R**, a T (number of time points) × P (number of voxels) matrix, contain the aspect of the data unexplained by the model. From these errors we can estimate the $P \times P$ variance-covariance matrix Σ:

$$\Sigma = \frac{1}{T}\mathbf{R}^T\mathbf{R}. \tag{4}$$

One option is to normalize each voxel by the standard deviation (σ_p) of its residuals, i.e. the square root of the diagonal of Σ.

$$b^+_{k,p} = \frac{b_{k,p}}{\sigma_p} \tag{5}$$

This means that response estimates in $\mathbf{b}_k$ from noisier voxels will be down-weighted. The same aim is achieved by using t values instead of regression estimates, which has been shown to increase classification performance of linear support vector machines (Misaki et al., 2010).

The second option is to not only suppress voxels with high error variance, but also to take into account the noise covariance between voxels. This leads to the multivariate extension of Eq. (5), which results in spatial pre-whitening of the regression coefficients:

$$\mathbf{b}^*_k = \mathbf{b}_k\Sigma^{-\frac{1}{2}}. \tag{6}$$

In this study, we estimate the covariance structure and apply noise normalization to each imaging run separately. One detail to consider is that the number of voxels may exceed the number of acquired volumes, which renders Σ rank-deficient and therefore non-invertible. To mend this, Σ is regularized by shrinking it towards the diagonal matrix, using the optimal shrinkage factor, i.e. the factor that minimizes the expected squared loss of the resultant covariance estimator (Ledoit and Wolf, 2004).

Multivariate noise normalization renders the noise component of the voxel response patterns approximately independent and identically distributed. Note, however, that spatial correlations due to voxel-by-voxel correlations in the true signal (Diedrichsen et al., 2011) will not be removed — hence noise-normalized patterns may still show considerable correlation structure.

Note further that computing the squared Euclidean distance on multivariately noise-normalized response patterns results in the squared Mahalanobis distance:

$$\begin{aligned} d^2_{Euclidean}\left(\mathbf{b}^*_k, \mathbf{b}^*_j\right) &= \left(\mathbf{b}^*_j - \mathbf{b}^*_k\right)\left(\mathbf{b}^*_j - \mathbf{b}^*_k\right)^T \\ &= \left(\mathbf{b}_j\Sigma^{-\frac{1}{2}} - \mathbf{b}_k\Sigma^{-\frac{1}{2}}\right)\left(\mathbf{b}_j\Sigma^{-\frac{1}{2}} - \mathbf{b}_k\Sigma^{-\frac{1}{2}}\right)^T \\ &= (\mathbf{b}_j - \mathbf{b}_k)\Sigma^{-1}(\mathbf{b}_j - \mathbf{b}_k)^T \\ &= \mathbf{c}\mathbf{B}\Sigma^{-1}\mathbf{B}^T\mathbf{c}^T \\ &= d^2_{Mahalanobis}(\mathbf{b}_k, \mathbf{b}_j). \end{aligned} \tag{7}$$

In the analyses presented here, we computed the Euclidean and the correlation distance on unnormalized, univariately (Eq. (5)), and multivariately (Eq. (6)) noise-normalized response estimates. Some methods (LDA, LDC, LDt; see Crossvalidation section and Pattern classifiers section) include multivariate noise normalization implicitly.

Crossvalidation

A problem for estimating distances from noisy data is that even if two patterns are in truth identical, the distance between the estimated patterns will be larger than zero, because noise makes the pattern estimates dissimilar.

To illustrate this, we simulated multiple instantiations of two random patterns with a true squared Euclidean distance ranging from zero to two. In each instantiation, we added varying degrees of i.i.d. noise to the patterns. We then calculated the squared Euclidean distance of these noisy patterns (Fig. 2A).

For very low levels of noise, the observed distances reflected the true distances accurately. For increasing levels of noise, however, the distance estimates increased independent of the true distance between conditions. Therefore, though the rank-order of distances can be interpreted, a value of zero and hence the ratio between distances is not meaningfully defined.

As a remedy, it has been suggested to split the data into independent partitions A and B and to validate the difference between k and j across them (Allefeld and Haynes, 2014; Kriegeskorte et al., 2007; Nili et al., 2014):

$$d^2_{Euclidean,crossvalidated}(\mathbf{b}_k, \mathbf{b}_j) = (\mathbf{b}_j - \mathbf{b}_k)_A(\mathbf{b}_j - \mathbf{b}_k)^T_B = \mathbf{c}\mathbf{B}_A\mathbf{B}^T_B\mathbf{c}^T. \tag{8}$$

Because noise is independent between A and B, the expected value of this estimate is zero if there is no systematic difference between the patterns for condition k and j. This is because the measured difference vectors ($\mathbf{b}_k - \mathbf{b}_j$) will point in random directions for each partition and will thus be close to orthogonal in a high-dimensional space.

Crossvalidated estimates of the distance (Fig. 2B) therefore do not grow with increasing noise, and their expected value reflects the true distance between patterns. This endows the distance estimate with a meaningful zero point, enabling us to statistically test whether two patterns show significant differences. Furthermore, the distance estimates now faithfully reflect the underlying distance structure, allowing us, for example, to test the hypothesis that one distance is twice as big as another distance. Such a test would be meaningless on non-crossvalidated distances, as the answer would largely depend on the noise level.

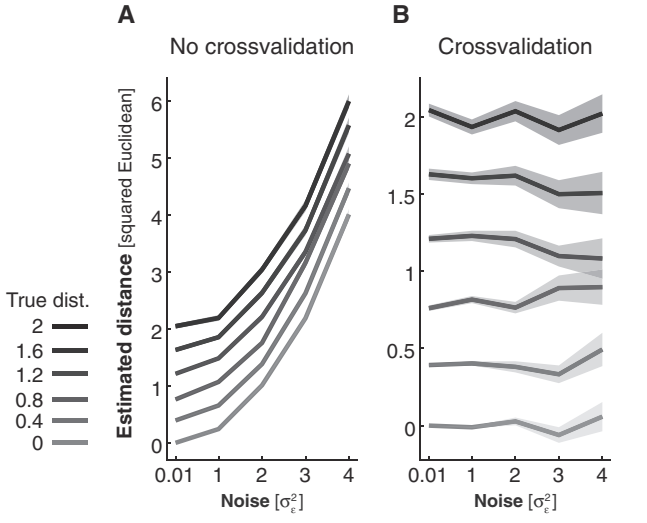

Fig. 2. Crossvalidation prevents the inflation of distance estimates by noise. Each line shows the average estimate squared Euclidean distance for a true squared Euclidean distance (ranging from zero to two) for different noise levels. Shaded error bars indicate standard error of the mean of 100 samples. (A) The non-crossvalidated distance estimate grows with increasing noise. (B) The crossvalidated distance estimate is robust against noise inflation.

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

Crossvalidation can also be applied to multivariately noise-normalized data, resulting in a crossvalidated estimate of the Mahalanobis distance (Eq. (9)). For reasons explained in the next section, we term this distance estimate *linear discriminant contrast* (*LDC*), as it closely relates to standard linear discriminant analysis (see Pattern classifiers section):

$$\begin{aligned} d^2_{Mahalanobis,crossvalidated}(\mathbf{b}_k,\mathbf{b}_j) &= (\mathbf{b}_j-\mathbf{b}_k)_A\Sigma_A^{-1}(\mathbf{b}_j-\mathbf{b}_k)_B^T \\ &= \mathbf{c}\mathbf{B}_A\Sigma_A^{-1}\mathbf{B}_B^T\mathbf{c}^T \\ &= LDC(\mathbf{b}_k,\mathbf{b}_j). \end{aligned} \quad (9)$$

Moreover, it has been suggested to normalize the LDC by an estimate of its standard error (Kriegeskorte et al., 2007; Nili et al., 2014). The resulting *linear discriminant t value* (*LDt*) can be used as an inferential measure of stimulus dissimilarity (for further details, see Appendix).

In this paper, we estimate the crossvalidated measures (LDC and LDt) in a leave-one-run-out crossvalidation, where one run was assigned to dataset A, and the remaining runs to dataset B. The distance estimates are then averaged across all possible crossvalidation folds.[1]

Pattern classifiers

Instead of directly estimating a distance measure between patterns, a number of fMRI studies have used pairwise classification accuracy as a proxy for pattern dissimilarity (e.g. Haxby et al., 2011, 2014; O'Toole et al., 2005; Pereira et al., 2009). Here, chance performance of classification corresponds to a zero distance.

One widely used classification approach is linear discriminant analysis, LDA (Fisher, 1936). LDA estimates a linear classification boundary under the assumption that the vectors $\mathbf{b}_k$ and $\mathbf{b}_j$ have a multivariate Gaussian distribution with separate true *1 x P* mean vectors and the same *P x P* within-class variance-covariance matrix Σ.

The Fisher discriminant rule is

$$v = \mathbf{w}\mathbf{b}^T \quad (10)$$

where

$$\mathbf{w} = (\mathbf{b}_j-\mathbf{b}_k)_A\Sigma_A^{-1}. \quad (11)$$

is the $1 \times P$ weight vector determining optimal classification. If v is larger than a criterion value c, the observation is assigned to class k, otherwise to class j.

If the test dataset B only consists of one observation of the two classes, and we subtract the mean pattern from both the training and the test dataset, then both observations will be correctly classified if

$$(\mathbf{b}_j-\mathbf{b}_k)_A\Sigma_A^{-1}(\mathbf{b}_j-\mathbf{b}_k)_B^T>0 \quad (12)$$

and incorrectly if this value is negative. Note that the classification function (Eq. (12)) is equivalent to the crossvalidated Mahalanobis distance, LDC (see Crossvalidation section). However, in LDA the discriminant is only used to make a binary decision for each response pattern, which then is converted into a classification accuracy. Therefore, the linear discriminant classification accuracy is tantamount to a discretized conversion of the LDC.

Another popular class of classification algorithms in fMRI are support vector machines (Ben-Hur et al., 2008; Cox and Savoy, 2003; Vapnik, 1995). Like LDA, SVM constructs a decision boundary between two classes. While the decision hyperplane can also be non-linear, it appears that the linear form yields higher performance in fMRI (Misaki et al., 2010). Unlike LDA, linear SVMs determine the classification boundary by maximizing the margin between the hyperplane and the closest training point on either side of it. This ensures that both classes are separated with maximum clearance. Like LDA, the SVM discriminant determines the class assignment whose accuracy is indicated by a percentage value. In this study, we used the LIB-SVM library (Chang and Lin, 2011) to perform the SVM analyses.

Like for the crossvalidated distances, classification accuracies were computed using a leave-one-run-out crossvalidation scheme, in which in each crossvalidation fold the classifier was trained on the data from all but one run, and then tested on the data from the remaining run. Classification accuracies were then averaged across crossvalidation folds. Before submitting the response patterns to the classification routine, we performed mean pattern subtraction for each run, which slightly increased classification accuracy, as it removes potential shifts of the whole pattern ensemble across imaging runs (dataset 1: +4.58%, 2: +4.45%, 3: +0.17%, 4: +0.16%).

RDM reliability analysis

A key requirement of a good dissimilarity measure is that it is reliable. Depending on the conclusions we wish to draw, however, the measure should replicate well on an ordinal scale (with preserved ranks), interval scale, ratio scale, or even in terms of its absolute magnitude.

We assessed reliability using split-half reliability estimates. To this end, we divided the data into two independent splits of odd and even runs: four runs per split for dataset one, two, and four; three runs per split for dataset three. The dissimilarity measures were then computed in each split. For the Euclidean and the correlation distance, we averaged the fMRI response patterns of each condition over runs before computing the distances. For the crossvalidated measures, we performed leave-one-run-out crossvalidation within each half of the data (see Crossvalidation section and Pattern classifiers section). Ultimately, we obtained two $1 \times Q$ vectors of dissimilarities, $\mathbf{m}_1$ and $\mathbf{m}_2$ (corresponding to split one and two), where $Q = K(K-1)/2$ pairwise distances for K conditions.

We computed four measures of RDM reliability: Spearman correlation, Pearson correlation, Pearson correlation with fixed intercept, and one minus the proportion of residual sum-of-squares. The Spearman correlation measures the correspondence between the RDMs in terms of their ranks, i.e. on an ordinal scale. The Pearson correlation assesses the stability of the relationship on an interval scale. However, both measures are mean-centered and therefore do not penalize any offset in the average distance across the two halves. Therefore, they do not provide information as to whether ratios of distances remain stable.

To assess their reliability, we computed a Pearson correlation that does not mean-center the values. Unlike the Pearson correlation, this measure is therefore not shift-invariant, but "fixes" the intercept of the regression line between the RDMs to zero. Finally, we computed the sum of squared differences between the Q distances from each split and divided them by the overall sums-of-squares of $\mathbf{m}_1$ and $\mathbf{m}_2$:

$$1-\frac{\sqrt{\sum_{q=1}^{Q}\left(m_{1_q}-m_{2_q}\right)^2}}{\sqrt{\sum_{q=1}^{Q}\left(m_{1_q}^2+m_{2_q}^2\right)}}. \quad (13)$$

Any difference between $\mathbf{m}_1$ and $\mathbf{m}_2$, may it be scaling or constant offset, will therefore reduce this reliability measure.

The last two reliability measures are only meaningful if the distance measure has an interpretable zero point, as they can change dramatically with an added constant value. We therefore only applied them to the crossvalidated measures (LDC, LDt, LDA, and SVM).

RDMs and split-half reliability measures were computed for each region-of-interest (ROI) of each subject (see fMRI data section). We

[1] This procedure lead to exhaustive crossvalidation, i.e. all possible crossvalidation folds are used. Other exhaustive schemes, such as considering all possible half-splits or all possible pairs of individual runs, will yield identical results.

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

then compared the RDM reliability across the four datasets using paired t tests with false discovery rate (FDR) at 5%.

fMRI data

We used four datasets from three independent fMRI experiments for the RDM reliability analysis. All experiments differed considerably with regards to paradigm (number and type of stimuli, number and length of trials per stimulus, number and length of baseline trials), data acquisition (number of subjects, number of functional runs, number of scanning sessions), scanning parameters (TR, volumes per run, voxel size), and functional regions considered (visual or motor areas, number of voxels included, see Fig. 3).

Dataset 1 & 2: Contralateral and ipsilateral finger representations in the primary motor and sensory cortex

The full study is described in Diedrichsen et al. (2013). Six participants underwent scanning while performing unimanual finger presses with the left and right hand. Finger presses were executed against an MRI-compatible keyboard and measured by a force transducer mounted underneath each key. Imaging data were acquired on a 3 T Siemens Trio with a 32-channel head coil. Eight functional runs of 126 volumes each using a 2D echo-planar imaging sequence (TR = 2.72 s) at voxel size 2.3 × 2.3 × 2.3 mm were recorded for each participant. Each trial was 8.16 s long, and each of the ten fingers was probed three times per run, resulting in 30 trials per run. The sequence of the fingers was fully randomized. In addition to these task-related trials, each run contained five randomly placed baseline trials of five or six TR length during which the subject was asked to rest. Anatomical ROIs were defined based on the probabilistic cytoarchitectonic maps aligned to the subject-average cortex surface reconstruction generated using Freesurfer (Dale et al., 1999; Fischl et al., 2008). Here, we use two ROIs which carry the most prominent neural representations of individual finger movements: the primary motor cortex M1 (Brodmann area 4) and the primary sensory cortex S1 (Brodmann areas 3a, 3b, 1, and 2). The average number of voxels across subjects was 757 (standard deviation 36 voxels) in M1, and 1492 (standard deviation 132 voxels) in S1.

While finger representations in M1 and S1 are predominantly activated during movements of the contralateral hand, they can also be activated to a lesser degree during movement of the mirror-symmetric fingers on the ipsilateral hand (Diedrichsen et al., 2013). Ipsilateral finger representations have a lower signal-to-noise ratio than their contralateral equivalents. We exploited this for our comparison by dividing the data into a contralateral and an ipsilateral dataset. We then evaluated a total of 12 hemispheres (six subjects × two hemispheres) for each contralateral (dataset 1) and ipsilateral (dataset 2) digit representation.

Dataset 3: Representations of visual objects at varying orientations

In this so far unpublished experiment, ten participants were presented with a total of 72 unique images of real-world objects. Each image belonged to one of three categories, namely faces, places, and leaves. From each category, two exemplars were displayed at 12 different orientations (0°, 30°, 60°, …,330°). All stimuli were grayscale, histogram-equalized, confined to a circular aperture, and presented at a retinal size of ten degrees visual angle. Before scanning, participants were familiarized with the stimuli. They learned to assign each exemplar to a predefined group (either A or B). The A/B labels were learned for the upright orientation of faces and places. For leaves, a random orientation was chosen for the learning phase. Imaging data were acquired on a 3 T Siemens Trio with a 12-channel head coil. Six functional runs containing 312 volumes each were measured using a 2D echo-planar sequence (TR = 2 s) with a voxel size of 3 × 3 × 3.75 mm. In each run, all 72 stimuli were presented twice in a random sequence, with a total of 144 trials per run. Each trial was 4 s long (SOA = 4 s). In each trial the image was displayed for one second and a gray background was presented for three seconds. Each one second presentation consisted of an image being flashed ON-OFF-ON-OFF-ON, where ON corresponds to the presentation of the image for 200 ms and OFF corresponds to the presentation of the gray background for 200 ms. In each trial subjects saw three flashes of the same image and were asked to respond during or after the presentation of the images if the displayed image was A or B. Additionally, each run contained 48 baseline trials (1 TR each) in which only a fixation cross was shown.

Functional ROIs were defined on independent data from a functional localizer experiment. The localizer images were recorded in one functional run of 203 volumes at TR = 2 s. The experiment contained images of four categories, faces, places, objects, and scrambled objects. Categories were presented at random in three blocks of 36 images each. Each image came with a superimposed fixation cross. While undergoing scanning, participants were asked to fixate and perform a one-back task. Two functional ROIs were defined in each hemisphere by their respective contrast: the fusiform face area, FFA (faces > places; Kanwisher et al., 1997) and the parahippocampal place area, PPA (places > faces; Epstein and Kanwisher, 1998). Both ROIs contained 84 voxels. Additionally we defined the human inferior-temporal cortex in each hemisphere by drawing an anatomical mask on the group-average cortical surface and backprojecting it into the single subject volume. This ROI comprised the 183 most responsive (by the contrasts all stimuli > baseline) voxels within the mask.

Dataset 4: The effect of categorization on visual object representations

In this as yet unpublished study, 17 participants underwent scanning in two separate sessions, each with four functional runs of 96 volumes. In each run, participants were presented with 24 images of real-

	Data set			
	1	2	3	4
Study type	motor	motor	object vision	object vision
Sample size	12	12	10	17
Number of ROIs	2	2	3	3
ROI (functional contrast or anatomical mask)	M1/S1	M1/S1	FFA/PPA/IT	FFA/PPA/LOC
Corresponding average ROI sizes	1,492/757	1,492/757	84/84/183	232/266/372
Number of timepoints per run	123	123	304	96
Number of runs	8	8	6	8
Number of conditions	5	5	72	24
Number of trials per condition	3	3	2	2
Brief description	contralateral finger presses	ipsilateral finger presses	visual presentation of faces, places & leaves	visual presentation of animate & inanimate objects

Fig. 3. The four fMRI datasets used in the RDM reliability analysis. Four fMRI datasets were analyzed: two condition-sparse (5 conditions, set 1 and 2) and two condition-rich sets (72 and 24 conditions, set 3 and 4). The condition-sparse datasets came from a motor experiment (set 1 and 2), while the condition-rich datasets were experiments on object vision (3 and 4).

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

world objects belonging to two categories (animate and inanimate) with 12 stimuli each. Stimuli were presented on a gray background at ~5 degrees visual angle (depending on the exact shape). A run contained 48 stimulus trials during which one of the images was flashed three times in a 500 ms ON, 500 ms OFF sequence. Each stimulus was presented during two trials and each trial lasted 3 s (SOA 3 s). In addition, each run contained 12 baseline trials during which only the gray background and a fixation cross were presented for 3 s. Trial order was randomized in each run. Participants were instructed to either categorize a stimulus based on the one previously shown (session one) or to complete a visual fixation task (session two). We used data from both sessions, yielding eight runs per subject. Each session also included a functional localizer of two runs during which participants viewed blocks of images depicting faces, houses, objects, and scrambled objects. Three functional regions were defined whose size varied between subjects: FFA (mean: 232; standard deviation: 47), PPA (mean: 266; standard deviation: 66), and the lateral occipital complex, LOC (mean: 372; standard deviation: 61), defined by the functional contrast objects > scrambled objects (Grill-Spector et al., 2001).

Functional EPI images covering the entire brain were acquired on a 3 T Siemens Trio scanner using a 32-channel head coil (2D echo-planar sequence, 32 slices, 3 mm isotropic resolution, inter-slice gap = 0.75 mm, TR = 2 s). For each participant we also obtained a high-resolution (1 mm isotropic) T1-weighted anatomical image using an MPRAGE sequence.

fMRI simulations

To confirm our empirical results, we also generated artificial fMRI data with a range of known signal-to-noise ratios (SNR). We simulated fMRI patterns for one condition-sparse and one condition-rich design. In the condition-sparse design, the number of conditions (5), trials (3), subjects (6), functional runs (8), time points per run (123), and the experimental design of the simulation were matched to dataset one and two. In the condition-rich design, the number of conditions (72), trials (2), subjects (10), functional runs (6), time points per run (304), and the experimental design corresponded to dataset three (see Dataset 3: Representations of visual objects at varying orientations question). We simulated fMRI regression coefficients for the P voxels of one ROI by drawing random $K \times 1$ vectors (K being the number of conditions) from a multivariate Gaussian with mean zero and variance-covariance matrix **G**. **G** determined the true similarity structure between the experimental conditions. The coefficients were then assembled in the $K \times P$ matrix $\mathbf{B}_{true}$. To generate fMRI timecourse data, we multiplied $\mathbf{B}_{true}$ with a design matrix **X** and added random Gaussian noise. We then step-wise increased the noise variance from 2 to 1000 times the signal variance. In the first sets of simulations the number of voxels was fixed to P = 123, in a third set we varied the numbers of voxels in the ROI across a range from 33 to 1419 voxels. In this simulation the noise level was adjusted such that the reliability of distances based on univariate noise normalization remained approximately constant.

In all simulations, noise was correlated across neighboring voxels, which is important to assess the performance of multivariate noise normalization under realistic conditions. The correlation between voxels *i* and *j* depended on their Euclidean distances $\delta_{i,j}$ and fell off as a Gaussian kernel with standard deviation s:

$$corr(i,j) = \exp\left(-\frac{\delta_{i,j}}{2s^2}\right) \quad (14)$$

For large values of s, the noise of neighboring voxels becomes highly correlated, for small values of s neighboring voxels become independent. In our simulations, s equaled 0.9.

The simulated fMRI patterns were submitted to the same analysis pipeline as the experimental data. Each simulation was repeated 1000 times and results were averaged across repetitions.

Results

The RDM split-half reliability scores and corresponding inference results are presented in Figs. 4 and 5 for all datasets. The datasets are sorted by their average RDM reliability (Pearson split-half correlation), from highest in dataset one to lowest in dataset four. Fig. 6 shows the results for the simulated fMRI datasets.

All distances and classifiers were applied to the response patterns after no, univariate or multivariate noise normalization. In summary, our results suggest that a) multivariate noise normalization improves the reliability of all dissimilarity measures; b) Euclidean and correlation distance are not significantly different in RDM reliability. However, the presence of category-selective univariate activation, the correlation distance tends to be numerically more reliable; c) crossvalidated distances do not lead to decreased reliability as compared to their non-crossvalidated counterparts; d) discretized classification accuracies are a significantly less reliable dissimilarity measure than continuous distances.

Multivariate noise normalization enhances the reliability of the dissimilarity measures

To statistically assess the influence of univariate and multivariate noise normalization, we pooled reliability scores across the Euclidean and the correlation distances.

Euclidean and correlation distance RDMs computed after univariate noise normalization produced significantly higher RDM reliability than their unnormalized counterparts in two out of four datasets for both Spearman and Pearson split-half correlations (Fig. 5, row "Noise normalization (univ. vs. none)"). This shows that accounting for noise contributions of individual voxels already has a positive effect on the distance estimates.

Multivariate noise normalization also takes into account the multivariate noise structure by down-weighting voxels with high noise correlations. Compared to the univariate noise normalization, multivariate normalization of the activation patterns always resulted in numerically higher reliability scores in the real fMRI data (Fig. 5, row "Noise normalization (multiv. vs. univ.)"). The difference in RDM reliability was significant or near significant in almost all datasets at uncorrected thresholding ($p < 0.05$). One comparison survived the FDR correction. These findings were replicated by our simulations, in which multivariate noise normalization improved RDM reliability of all measures over univariately normalized patterns (Fig. 6A). This effect was present for both distance measures and classifiers as well as for both the condition-sparse and the condition-rich design, although more sustained in the former.

Together, these results clearly show that normalizing by the estimate of the full noise covariance Σ stabilizes the distance estimates more effectively than univariate normalization.

Optimal shrinkage safeguards the multivariate noise normalization

When multivariate noise normalization is applied to large ROIs, the number of voxels can easily be higher than the number of time points (e.g. see the section Dataset 1 & 2: Contralateral and ipsilateral finger representations in the

primary motor and sensory cortex), resulting in a rank-deficient estimate of Σ. To attain invertibility, we used optimal shrinkage of Σ towards a diagonal noise matrix (Ledoit and Wolf, 2004). With increasing number of voxels the shrinkage algorithm will regularize the noise covariance matrix more, and in the extreme case will converge to a diagonal covariance matrix, thus turning multivariate into univariate noise normalization.

Across experiments the average shrinkage was between 6%–16%, with the highest values for experiment 1 and 2, in which the number

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

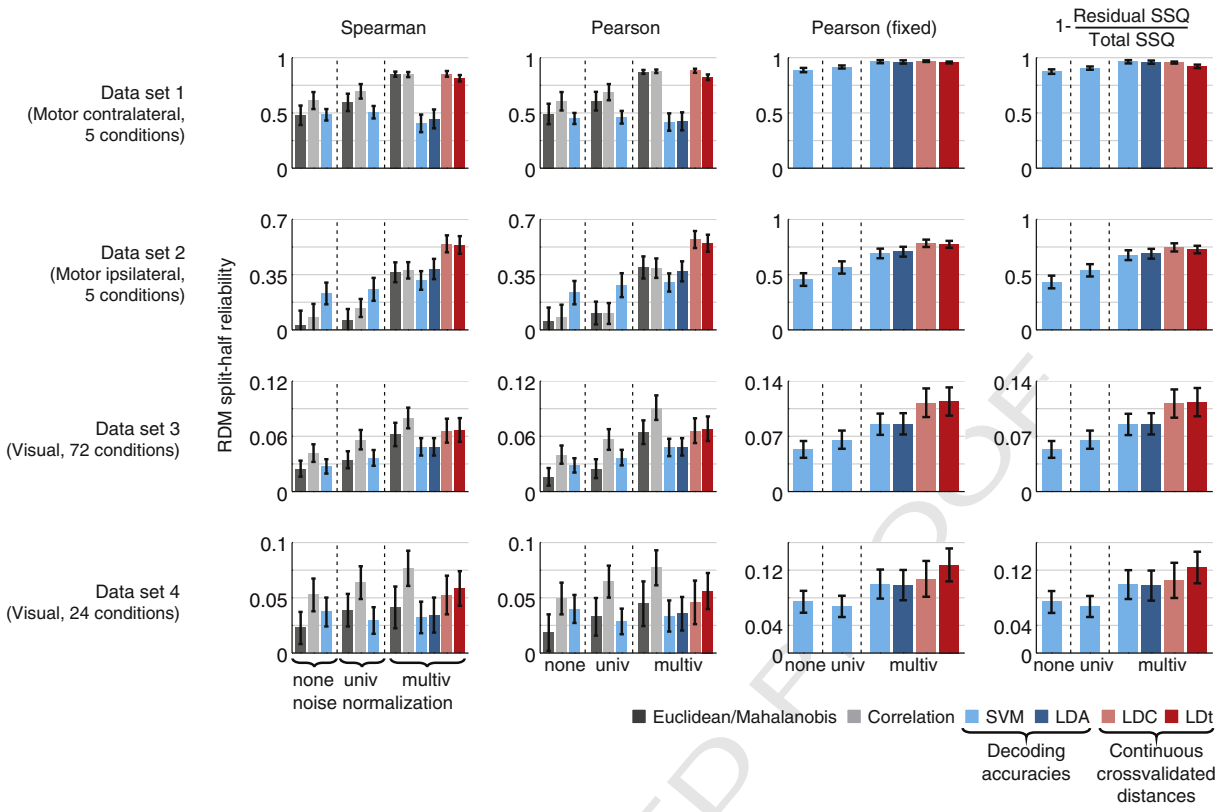

Fig. 4. RDM split-half reliability analysis of four fMRI datasets. We assessed the RDM split-half reliability of all dissimilarity measures using Spearman correlation, Pearson correlation, Pearson correlation with fixed intercept, and the residual sum-of-squares (see RDM reliability analysis section). The latter two measures were only applied to the crossvalidated dissimilarity measures (LDC, LDt, LDA, and SVM). The bar graphs show the RDM reliability scores of the dissimilarity measures using no normalization (none), univariate, and multivariate normalization. Error bars indicated standard errors across subjects and ROIs. Note that the y-axes are on different scales for different datasets.

	Data set *(N)*			
Comparison *(reliability measure)*	**1** *(12)*	**2** *(12)*	**3** *(10)*	**4** *(17)*
Spearman RDM reliability				
Noise normalization (univ. vs. none)	**0.001**	0.428	**0.005**	0.115
Noise normalization (multiv. vs. univ.)	0.058	**0.037**	**0.031**	0.312
Correlation distance vs. Euclidean (multiv.)	0.857	0.817	0.176	0.11
LDC vs. Mahalanobis	0.885	**0.03**	0.739	0.657
Distance (LDC & LDt) vs. accuracy (LDA & SVM)	**0.03**	**0.002**	**<0.001**	**0.017**
Pearson RDM reliability				
Noise normalization (univ. vs. none)	**0.002**	0.315	**0.004**	0.109
Noise normalization (multiv. vs. univ.)	0.056	**0.042**	**0.007**	0.166
Correlation distance vs. Euclidean (multiv.)	0.623	0.942	0.051	0.162
LDC vs. Mahalanobis	0.365	**0.031**	0.898	0.961
Distance (LDC & LDt) vs. accuracy (LDA & SVM)	**0.029**	**0.001**	**<0.001**	0.113
Pearson RDM reliability (fixed intercept)				
Distance (LDC & LDt) vs. accuracy (LDA & SVM)	0.984	**0.014**	**<0.001**	0.085
1- Residual SSQ / Total SSQ				
Distance (LDC & LDt) vs. accuracy (LDA & SVM)	0.163	0.092	**<0.001**	0.103

FDR < 5% p < 0.05 (uncorrected)

Fig. 5. Crossvalidated continuous distance estimates using multivariate noise normalization are most reliable. The table shows p values for comparisons of the RDM reliability measures (Fig. 4). Each row lists a comparison for a given RDM reliability measure. Each column lists one of the four fMRI dataset. Values were computed using paired *t* tests. Light red: significantly greater RDM reliability at an uncorrected threshold of $p < 0.05$. Dark red: significant after an FDR correction at 5%.

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

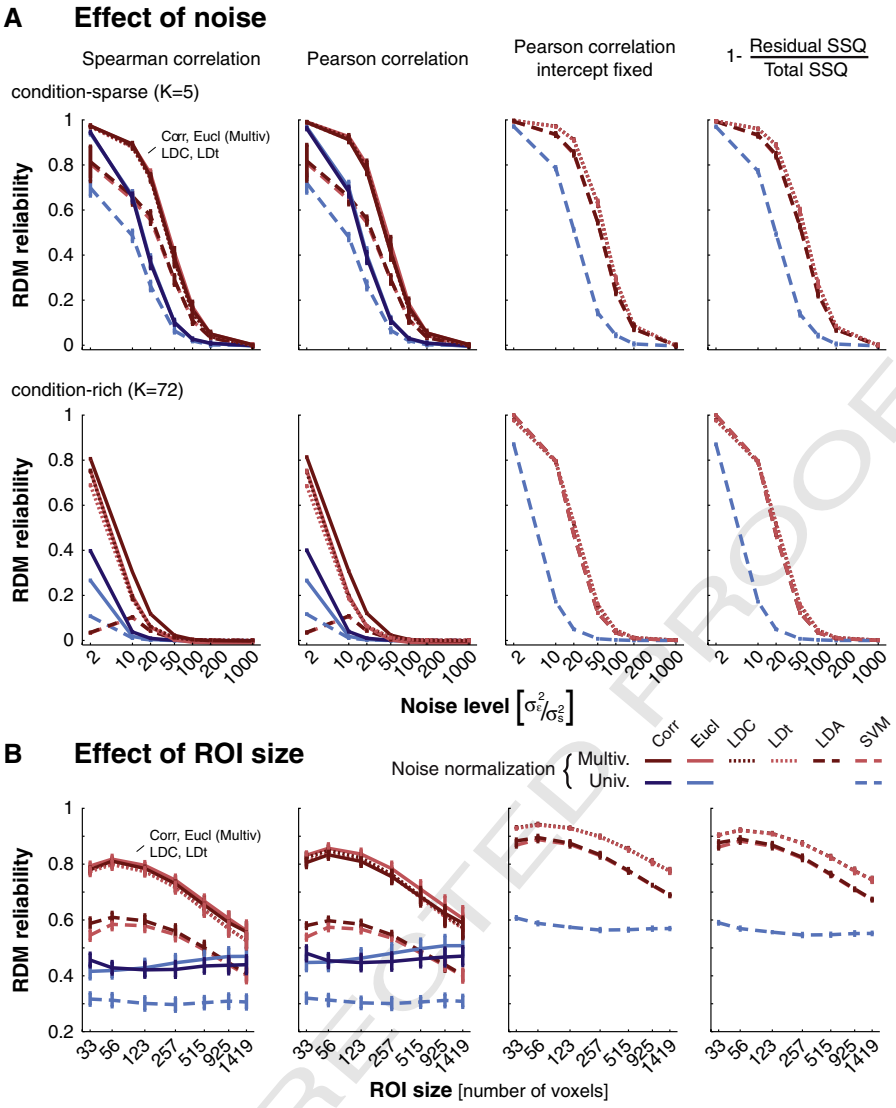

Fig. 6. RDM reliability analysis of simulated fMRI data. The graphs show the average RDM reliability scores from 1000 simulated experiments. Simulations were carried out for both univariate (blue lines) and multivariate (red lines) noise normalization, and all dissimilarity measures: distances (solid lines, Euclidean and correlation distance), crossvalidated distance estimates (dotted lines, LDC and LDt); and classification accuracies (dashed lines, LDA and SVM). Error bars indicate the standard error for the simulated sample size. (A) Simulation results for a condition-sparse design (five conditions, six subjects) and a condition-rich design (72 conditions, ten subjects) at varying levels of noise. Continuous distance measures with multivariate noise normalization perform best. Classification accuracies are less reliable than distance. (B) Simulation results for varying ROI size. For larger ROIs, the noise was increased such that the split-half reliability of univariate noise normalization was approximately constant. Multivariate noise normalization leads to higher RDM reliability even at large ROI sizes.

of voxels outstripped the number of volumes by factor 10:1. Even in these cases, however, multivariate noise normalization had a clear advantage over univariate noise normalization (see Fig. 4, dataset 1 and 2).

To further investigate the effect of ROI size on multivariate noise normalization, we simulated fMRI patterns with a varying number of voxels (Fig. 6B). The design of the simulation was identical to dataset 1 and 2 (see fMRI simulations section). To keep the amount of signal constant across ROI sizes, we scaled the true pattern variance-covariance matrix **G** by $P^{-0.45}$ (P being the number of voxels). With increasing ROI size, RDM reliability of multivariately normalized response patterns approached the performance of univariate noise normalization as a result of shrinkage. However, multivariate noise normalization yielded robustly higher average RDM reliability even at a very large ROI size (>1000 voxels). This shows that multivariate noise normalization can be applied even when the number of voxels drastically outnumbers the number of time points.

Euclidean and correlation distance are similarly reliable

We then compared the reliability of the Euclidean and correlation distance. When using either none or univariate noise normalization, RDM reliability scores of the Euclidean and correlation distances were tightly matched. We only found a significant advantage of the correlation distance in dataset one when patterns had not been noise-normalized (Spearman RDM reliability: $t_{11} = 4.45$, $p = 0.001$; Pearson RDM reliability: $t_{11} = 3.28$, $p = 0.008$). We found no significant difference between the distances for univariate normalization.

Employing multivariate normalization, Pearson and Spearman RDM reliabilities of the Euclidean and correlation distance were again very similar (Fig. 4), with no significant difference between the two (Fig. 5, row "Correlation vs. Euclidean (multiv.)"). Moreover, we found no difference between the distance measures in the condition-sparse simulation (Fig. 6A, top row). However, we observed higher RDM reliability of the correlation distance up to intermediate noise levels in the simulated

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

condition-rich design (Fig. 6A, bottom row). This difference in reliability is likely a result of the categorical structure in dataset three (see The correlation distance is sensitive to activation differences section).

Overall, we observed that correlation distance RDM reliability was numerically higher in most studies, but seldom significantly so. Moreover, the RDM structure of the correlation distance also reflects condition-specific activation, which is likely to be common to both splits and may increase reliability.

Crossvalidation improves, rather than impairs, RDM reliability

Contrary to conventional distance measures, crossvalidated distance estimates are unbiased by noise (Crossvalidation section). Therefore they can be statistically compared against zero to test whether the response patterns of two conditions are significantly different. Furthermore, increasing noise does not distort the structure of the representational space (Fig. 2).

These advantages come at the cost of having to split the data to allocate them to training and test sets. Although crossvalidation ultimately still uses all the data available, this splitting might decrease the RDM reliability. We therefore tested the performance of the Euclidean distance after multivariate noise normalization (i.e. the Mahalanobis distance) against LDC (i.e. the crossvalidated Mahalanobis distance). Contrary to our expectation, LDC was not significantly less reliable than the Mahalanobis distance in all datasets (Fig 5, row "LDC vs. Mahalanobis"). Quite the opposite, in datasets two and four LDC produced even more reliable RDMs (Spearman and Pearson split-half correlation). This result was also confirmed by the fMRI simulations, in which the Mahalanobis distance and LDC performed equally (Fig. 6A). Moreover, we found that LDt was slightly less reliable than other distance measures in the condition-rich design, but only when noise was extremely low; this is not due to crossvalidation, but because LDt does not scale linearly in the noise limit (see Eqs. (A4) and (A3) in the appendix). Overall, these results show that in the case of LDC the advantages of crossvalidated distance measures do not trade off against their reliability.

Continuous distance measures are a more reliable and more informative dissimilarity measure than classification accuracy

We now turn to the question of whether continuous distance measures or discretized classification accuracies are a more reliable measure of brain representations. To investigate this, we only consider the results for the multivariate noise normalization (which is implicit in LDA, LDC, and LDt). This allowed for a fair comparison, because classifiers and distance measures profited from the same noise normalization and used the same crossvalidation scheme (leave-one-run-out).

We found that for real fMRI data, RDM reliabilities were significantly higher for the distance estimates than for classification accuracies in most cases (Fig 5., row "Distance (LDC & LDt) vs. accuracy (LDA & SVM)"). This finding was replicated by the fMRI simulations, where RDMs based on continuous distance measures were consistently more reliable than those based on classification accuracies (Fig. 6A).

Why are linear classifiers less reliable estimators of representational geometry than continuous distance measures? As pointed out in the Pattern classifiers section, the LDA classifier is closely related to the more reliable crossvalidated LDC distance measure. However, there are three potential factors that may reduce the reliability of the classification accuracy measure.

First, classification accuracy is inherently bounded by 100%, whereas continuous distance measures can increase, even if the two patterns are already perfectly separated. This feature is the reason for the decreasing reliability in the simulations when noise levels are very low (Fig. 6A, dashed lines in Spearman and Pearson split-half correlation). In practice, however, this does not constitute a major problem, as classification accuracy is typically well below 100%.

Second and more importantly, classification accuracy is a measure obtained from binary decisions, which discard continuous dissimilarity information (see Pattern classifiers section). This lossy conversion alone could make the accuracy RDMs less reliable.

Finally, the decision criterion needs to also be learned from the training data, and is then applied to the test set. It has been shown that the average mean pattern varies considerably between imaging runs (Diedrichsen et al., 2011) and also slowly changes within each imaging run (Henriksson et al., 2015). Because the classification boundary is optimized for the training set, it is unable to cope with shifts of the pattern ensemble in the test set. This will reduce classification accuracies, but likely also result in less reliable RDMs.

To evaluate the effect of discretization and pattern shift on classification accuracy, we performed an fMRI simulation similar to the ones described in fMRI simulations section. We simulated fMRI response patterns of 10 conditions for 100 subjects with 20 runs each. We added an idiosyncratic mean-pattern to each run, leading to a shift of the pattern ensemble. We then varied the strength of the run-specific mean-pattern and compared the RDM split-half reliability (Pearson correlation) of LDA (discretized classification accuracy) to LDC (continuous distance). The LDA was performed on response patterns with or without prior mean pattern subtraction (see The effect of mean pattern subtraction (cocktail-blank removal) section).

RDM split-half reliabilities are shown in Fig. 7. First, the results confirm that the continuous distance estimate is more reliable than the classification accuracy, even in the absence of pattern shift — an advantage that is due to discretization. Secondly, as the pattern shift grew stronger, classification accuracy became less reliable. This is because the optimal classification boundary differed increasingly between training and test set. Third, we found that mean-pattern removal restored reliability of classification accuracy to baseline. Moreover, mean-pattern removal also increased the average classification accuracy from 13%

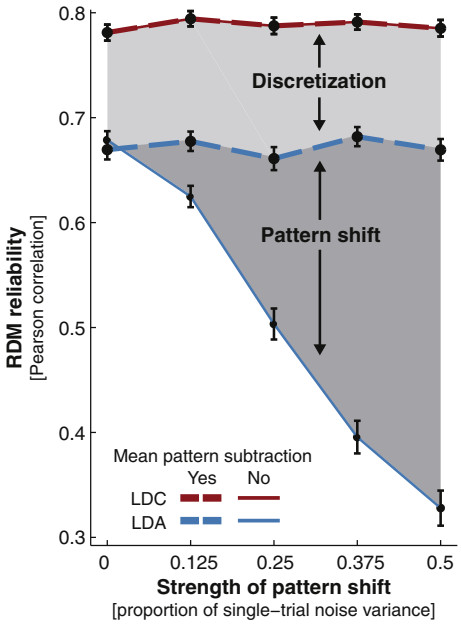

Fig. 7. Classification accuracies are less reliable due to discretization and mean-pattern shifts. The graph depicts the average RDM split-half reliability (measured as the Pearson split-half correlation of two independent data splits) of LDA and LDC for 100 simulated subjects (error bars show subject standard error). Each simulated run contained a unique mean pattern whose strength was gradually increased. RDMs based on a continuous dissimilarity measure (LDC) are consistently more reliable than those based on classification accuracy (LDA). For a pattern shift strength of 0, this difference is explained by the discretization implicit in the classification (see Pattern classifiers section). With increasing run-pattern strength, LDA reliability decreased, while LDC reliability remained unaffected. This effect could be eliminated through mean pattern subtraction.

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

(unnormalized) to 31% above chance level. Finally, continuous distances were not affected by the shift of the mean pattern, as they do not apply a decision criterion. Taken together, these results strongly favor the use of continuous distances over classifiers when investigating brain representations.

The correlation distance is sensitive to activation differences

The Euclidean and correlation distance express similarity in fundamentally different ways and are therefore susceptible to different sources of variability (Fig. 1). The following section shows how stimulus-related activation influences the distance measures. We will illustrate this property using two ROIs of dataset three, FFA and PPA (see Dataset 3: Representations of visual objects at varying orientations section). These ROIs are known to show strong face- and place-selective activation respectively, which was also confirmed in this study (Fig. 8A).

For both regions, we computed RDMs using the Euclidean and the correlation distance. We then determined the average within-category dissimilarity (Fig. 8B, bar graphs), which is defined as the average distance between all stimuli of the same category (here 24 stimuli for each of the three categories). While the within-category Euclidean distances were similar for faces, places and leaves, the average correlation distance was significantly lower for stimuli of the preferred category (red bars) compared to non-preferred categories (gray bars; faces in FFA: $t_9 = 12.337$, $p < 0.001$; places in PPA: $t_9 = 6.813$, $p < 0.001$). If one interpreted the within-category distances as a measure of the sensitivity with which this region represents small stimulus differences, the correlation distance would lead us to claim that FFA is especially insensitive (or invariant) to different orientations of faces.

Why is the correlation distance relatively small when stimulus-activation is high? The explanation is that in the *P*-dimensional voxel space, those patterns correspond to points that are moved away from the origin by the shared activation. As a result, the angles between the corresponding vectors will be small on average. Relative to that, pattern vectors associated with other conditions will be closer to the origin, with larger angles between them. Therefore, the correlation distance will be small for faces and large for non-face stimuli in FFA (the same applies to PPA for places). Such a prominent difference in stimulus activation also contributes to the RDM reliability of the correlation distance (see Figs. 4 and 6). In contrast, the Euclidean distance and derived methods (Mahalanobis distance, LDC, LDt) do not depend on the angle, but on the distance between the pattern vectors. Because all patterns are moved by a comparable amount by the common activation (Fig. 8A), Euclidean distances are not reduced for categories that lead to large activation.

Discussion

RSA has found widespread applications in neuroimaging. One crucial choice the investigator faces is which dissimilarity measure to use. Surprisingly, to date no systematic comparison about the reliability of dissimilarity measures has been published. The analyses performed in this study strongly suggest four conclusions.

(a) Activation patterns (usually formed by regression coefficients) should be subjected to multivariate noise normalization to improve RDM reliability, regardless of dissimilarity measure.
(b) Continuous distances are more reliable and informative than classification accuracies as the latter are compromised by a ceiling effect, discretization, and run-specific pattern shifts.
(c) The Euclidean/Mahalanobis distance and the correlation distance are similarly reliable. However, the correlation distance is harder to interpret because conditions eliciting little activity have essentially uncorrelated patterns and thus large correlation distances, even though the patterns may not be significantly different (Fig. 8). In other words, a correlation distance of 1 can indicate either statistically distinct patterns or identical patterns.
(d) Crossvalidated distance estimators are unbiased. They have a

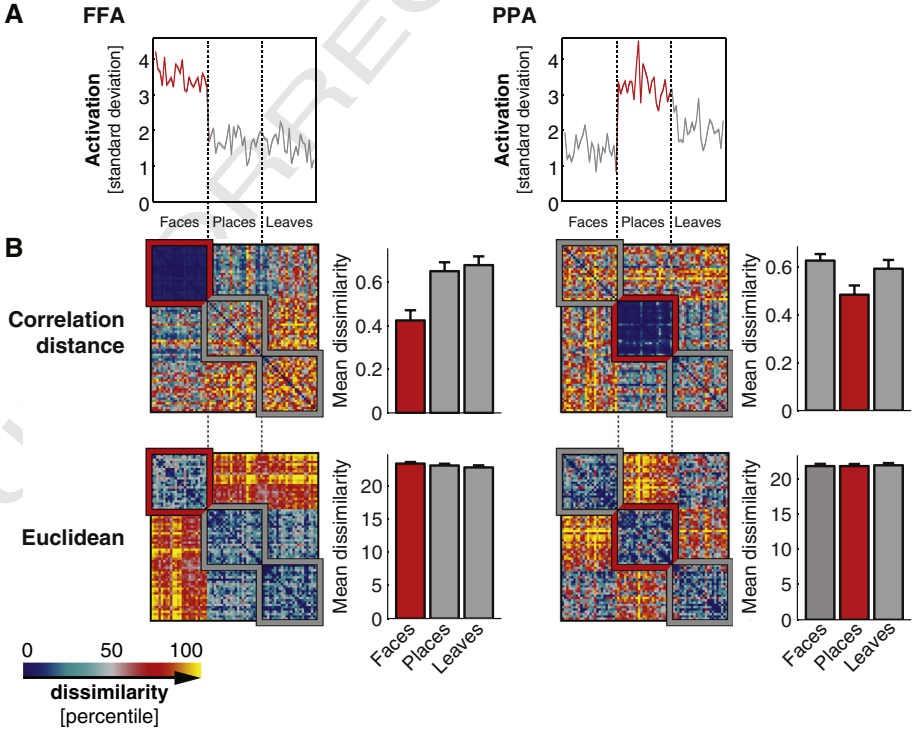

Fig. 8. The correlation distance is sensitive to differences in stimulus activation. Activation and RDM analysis of response patterns in FFA and PPA in dataset three (see the section Dataset 3: Representations of visual objects at varying orientations). The preferred stimulus category (faces for FFA, places for PPA) is highlighted in red. (A) Mean activation profile of the functional regions. As expected, both regions show higher activation for their preferred stimulus type. (B) RDMs and bar graphs of the average distance within each category (error bars indicate standard error across subjects).

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

meaningful zero point and enable ratios between distances to be interpreted.

As an overall conclusion, crossvalidated distance estimators with multivariate noise normalization are the method of choice when investigating brain representations with RSA.

Multivariate noise normalization: accounting for covariances in the fMRI noise leads to more reliable representations

The results of the RDM reliability analysis of three fMRI experiments and fMRI simulations (see Multivariate noise normalization enhances the reliability of the dissimilarity measures and Optimal shrinkage safeguards the multivariate noise normalization sections) convincingly demonstrate that multivariate noise normalization significantly improves RDM reliability. This improvement was observed for both continuous distances and classification accuracies. Misaki et al. (2010) have already shown that univariate noise normalization (using t values) results in higher classification accuracy compared to unnormalized regression coefficients. Here we show that noise normalization also leads to more replicable RDMs (regardless of accuracy level) and that even larger gains in reliability can be obtained when applying multivariate noise normalization.

Both real data and simulations show that the benefit of multivariate noise normalization is present across all noise levels except for very high (where the reliability is at floor) or very low noise (where reliability is at ceiling). Moreover, multivariate noise normalization is beneficial in both condition-sparse and condition-rich designs (Fig. 6A). We also found that the improvement in reliability was even present when the number of voxels outstripped the number of available data points by 10:1 (Fig. 6B). This is somewhat surprising, as the estimate of the variance-covariance matrix needs to be regularized when the number of voxels exceeds the number of time points, which can severely impair classification accuracy in LDA (Cox and Savoy, 2003) where the decision boundary depends on Σ. However, we found that multivariate noise normalization in conjunction with shrinkage worked well for large ROIs, even though our simulation indicated that the gains become somewhat smaller. For a large number of voxels, regularization biases the estimate of the covariance matrix towards a diagonal matrix and therefore makes multivariate noise normalization more similar to univariate noise normalization. Taken together, these results demonstrate that multivariate noise normalization can be effectively applied irrespective of the voxel-to-time point-ratio.

Multivariate noise normalization and the spatial scale of the fMRI signal

Dissimilarity estimates are more reliable after multivariate noise normalization, but are they also systematically different from dissimilarity estimates without noise normalization? The answer to this question is not straightforward and will depend on the spatial scale of the informative fMRI signals and the spatial scale of the noise processes. Multivariate noise normalization will de-emphasize voxels with correlated noise, and emphasize voxels that are uncorrelated. This can be understood as a form of spatial filtering. For example, if all voxels in an ROI are correlated equally strongly in the noise, multivariate noise normalization will remove the lowest spatial frequency (the mean) of the patterns. If neighboring voxels are more correlated with each other, multivariate noise normalization will remove the corresponding middle frequencies, thereby emphasizing the differences between immediately neighboring voxels (the highest spatial frequencies). Whether this spatial filter would systematically bias the RDM estimate depends on the spatial structure of the true signal. If the RDM structure is the same across all spatial scales (i.e. the RDM is the same no matter whether you look at high or low spatial frequencies), multivariate noise normalization will not bias the RDM estimate, but simply ensure the optimal (i.e. lowest variance) estimate. If the RDM structure changes with the spatial scale then one may find systematic differences between multivariate and univariate noise normalization. Consider an experiment presenting different exemplar of faces and scenes. Distances between faces and scenes may be mostly at low spatial frequencies, as these two items activate different ROIs (FFA and PPA, respectively) within the inferior-temporal cortex (IT). Distances between specific faces or scenes may rely on finer voxel-by-voxel differences within these regions and hence on higher spatial frequencies. For IT response patterns, multivariate noise normalization would likely render the between-category differences smaller and the within-category differences larger. Notwithstanding this feature, it should be noted that the inherent spatial resolution of fMRI already introduces an arbitrary choice regarding the spatial scale at which the RDMs are measured. Multivariate noise normalization simply biases the RDM to the spatial frequencies that is best measured with fMRI: usually towards slightly higher spatial frequencies than univariate noise normalization.

Classifiers vs. distances

Our results show that under equal conditions, continuous distance estimates provide a more reliable and nuanced dissimilarity measure than classification accuracies. It is important to note that the classifiers employed here also fundamentally rely on the notion of distance: LDA classifies test patterns according to their Mahalanobis distance from the class means (e.g. Bishop, 2006); SVM estimates the support vectors by maximizing the minimum distance to the boundaries of the training examples (Vapnik, 1995). However, both methods restrict themselves to estimating a percentage of all the test patterns that fall on the correct side of the decision boundary, i.e. they transform distances into binary yes-no decisions. By contrast, the continuous distances investigated here reflect the similarity of the stimulus patterns more directly, resulting in higher RDM reliability. The tight correspondence between these methods is especially obvious in the case of LDC, which utilizes the weight vector of LDA. For this reason, the results presented in this paper are likely to generalize to other measures of pattern discriminability that map a direct measure of similarity (classification weights) onto a less detailed scale (discrete percentages).

Previous studies have compared different classification methods for MVPA, recommending one over the other because it resulted in higher classification accuracy (Cox and Savoy, 2003; Grosenick et al., 2008; Ku et al., 2008; Misaki et al., 2010). However, most investigators are not interested in obtaining high classification accuracies, but rather in sensitively detecting whether a region encodes a certain variable or in determining whether one variable is more prominently encoded than another. For this purpose, a high reliability of the dissimilarity measure is much more important. In this respect, our results show that SVM and LDA classification fall short against continuous distance measures in almost all comparisons. This strongly suggests the use of distance measures over classifiers when investigating brain representations.

Crossvalidation

Crossvalidation of distance measures in fMRI has recently been proposed in the form of LDt (Kriegeskorte et al., 2007; Nili et al., 2014) and as part of a more general MANOVA framework (Allefeld and Haynes, 2014). We show here that the expected crossvalidated distance between two noisy estimates of the same pattern is zero, and that crossvalidated distance estimates are noise-unbiased. Moreover, we found that the crossvalidated Mahalanobis distance, LDC, was equally or even more reliable than its non-crossvalidated counterpart despite data splitting.

These features make crossvalidated distance estimates very attractive. First, crossvalidation enables us to infer whether the response patterns of two conditions are significantly different, by simply comparing the distances against zero. Therefore, crossvalidated distances can be

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

used in a similar fashion as classification accuracies, which are compared to chance performance. While inference is enabled by the fact that the expected value of a crossvalidated distance estimate is zero under the null hypothesis, we still need a measure of its distribution for statistical testing. This can be obtained, as with other multivariate methods, through permutation methods, i.e. by randomly exchanging condition labels and recalculating the dissimilarity measure (Kriegeskorte et al., 2006; Stelzer et al., 2013). In the context of multi-subject experiments, however, the distances can also be used as input to a traditional, parametric group analysis.

Second, the ratios or relative sizes of crossvalidated distances can be meaningfully interpreted, even across different regions or subject with different noise levels. This allows us to test richer and more detailed representational models than was possible when only considering the rank-order of distances (Kriegeskorte et al., 2008).

Overall, crossvalidated distance estimates are recommendable as they are not inflated by noise and endowed with an interpretable zero point. This has also implications for dimensionality-reduction algorithms for data visualization such as multidimensional scaling (Borg and Groenen, 2005) and *t* distributed stochastic neighbor embedding (Maaten and Hinton, 2008), where a meaningful zero value of the dissimilarity measure adds to the interpretability.

Euclidean vs. correlation distance

Though performing nearly equally in all cases, the correlation distance was oftentimes numerically slightly more reliable than the Euclidean distance. Partly this difference may be caused by the fact that the correlation distance reflects to some degree the size of the activation common to the different categories. Specifically, we showed that the correlation distance becomes smaller in the presence of a strong category-specific mean activation pattern, as observed in dataset three (Fig. 8). In such a case, much of the structure of the RDM will be influenced by the mean pattern activation, which may add to the reliability of the RDM, but may change its interpretation. Therefore, the choice between the distance measures depends strongly on the question that the investigator wants to answer. Oftentimes this will be how discriminable multiple stimuli are from each other. In this case, Euclidean-type distances like the Mahalanobis distance provide a good choice, as they are uninfluenced by the strength of a common activation pattern. By contrast, if one would like to establish how similar two response patterns are in terms of their specific shape, independent of the strength of the activation, correlation distances may be a good option.

Another advantage of Euclidean measures is that crossvalidation is easily achieved. While crossvalidated versions of correlation coefficients are possible, it is not straightforward to construct a correlation distance that is unbiased with respect to the noise in the same way as LDC (Fig. 2).

Conclusions

Across a range of datasets, we found that the crossvalidated Mahalanobis distance (LDC), which includes multivariate noise normalization, provides the most reliable measure of pattern dissimilarity. This measure combines the advantages of continuous distance measures and classification approaches. Like traditional distances, the measure is continuous, making it more reliable and informative. Like classification accuracy, it is crossvalidated, therefore unbiased, and can directly be used to test whether two response patterns are distinct. Finally, unlike any other approach, it provides ratio-scale representational dissimilarities, and thus a richer characterization of the representational geometry. These features make the crossvalidated Mahalanobis distance a powerful tool to investigate brain representations.

Acknowledgements

This project was supported by the Gates Cambridge Scholarship to AW, a European Research Council Starting Grant (ERC-2010-StG 261352) and a Wellcome Trust Project Grant (WT091540MA) to NK, and a Wellcome Trust Project Grant (094874/Z/10/Z) to JD. The Wellcome Trust Centre for Neuroimaging at UCL is supported by core funding from the Wellcome Trust (091593/Z/10/Z). The authors declare no conflict of interest.

Appendix

Linear discriminant t value

LDC is the difference between two conditions measured along a linear discriminant that has been estimated with independent data. In analogy to a univariate test, the LDC is a contrast measured on the discriminant. It generalizes the contrast measured for the average activation of an ROI to arbitrary weighted combinations of the ROI voxels (where the weights have been chosen with independent data to maximize sensitivity to the difference between the two conditions). Like any linear model contrast, the LDC can be converted to a *t* value by normalizing it by its standard error. We refer to this measure as the linear-discriminant *t* value (LDt; Nili et al., 2014; Kriegeskorte et al., 2007). The LDt is valid t value, which can be converted to a p value. An entire LDt RDM can be inferentially thresholded using the false-discovery rate (FDR), which is unaffected by the row- and column-wise dependencies of the LDt values.

To compute the standard error of the LDt, s_B, we estimate the error variance ${\sigma_{\epsilon_B}}^2$ on the residuals of the test set B and project it onto the discriminant $\mathbf{w}$ (Eq. (11)):

$$s_B = \sqrt{\left(\mathbf{w}\sigma^2_{\epsilon_B}\mathbf{w}^T\right) * \left(\mathbf{c}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{c}^T\right)} \tag{A1}$$

with c = [1–1]. LDt is then

$$LDt(\mathbf{b}_k, \mathbf{b}_j) = \frac{LDC(\mathbf{b}_k, \mathbf{b}_j)}{s_B}. \tag{A2}$$

The LDt is Student-*t* distributed under the null hypothesis that the two patterns are identical.

Note, however, that this feature is compromised when averaging LDt values. For example, when averaging LDt values across crossvalidation folds, the resulting average LDt is not t distributed. Furthermore, the LDt values from different folds are not independent, such that the standard error cannot be simply divided by the square root of the number of folds (Nili et al., 2014). As another example, it is often useful to average LDt values across different pairs of conditions, to assess within- or between-category information. The resulting average LDt values will again not be t distributed. However, as the LDC, the LDt measure is distributed symmetrically around zero under the null-hypothesis that the patterns are identical, and therefore can be used as a basis for other inference procedures, including condition-label randomization or bootstrap tests and for group-level inference with subject as random effect.

One potential drawback of LDt compared to LDC is that the relationship between the distances changes with the level of noise. To simplify the following illustration, assume the fMRI patterns have already been successfully multivariately noise-normalized; we can thus ignore the noise covariance matrix in the LDA weight vector. We can therefore rewrite the LDt as:

$$LDt(\mathbf{b}_k\mathbf{b}_j) = \frac{(\mathbf{b}_j - \mathbf{b}_k)_A (\mathbf{b}_j - \mathbf{b}_k)_B{}^T}{\sqrt{(\mathbf{b}_j - \mathbf{b}_k)_B (\mathbf{b}_j - \mathbf{b}_k)_B{}^T c}} \tag{A3}$$

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012

where c is a noise-dependent constant that is the same across any pair of conditions *j* and *k*, i.e. a constant that does not influence the ratios between different distances of a single ROI.

Because the regression coefficients **b** are estimates of the true response patterns β and are corrupted by noise, the expected value of the inner product in the denominator is

$$E\left(\left(\mathbf{b}_j-\mathbf{b}_k\right)_B\left(\mathbf{b}_j-\mathbf{b}_k\right)_B{}^T\right)=\left(\beta_j-\beta_k\right)\left(\beta_j-\beta_k\right)^T+K \tag{A4}$$

where *K* increases with the level of noise and is independent of the particular distance (assuming that all conditions are affected by equally high measurement noise). In contrast, the expected value of the inner product of the numerator (i.e. LDC) is crossvalidated and hence is independent of the noise:

$$E\left(\left(\mathbf{b}_j-\mathbf{b}_k\right)_A\left(\mathbf{b}_j-\mathbf{b}_k\right)_B{}^T\right)=\left(\beta_j-\beta_k\right)\left(\beta_j-\beta_k\right)^T. \tag{A5}$$

Hence, for high noise levels, the term *K*, which is the same across all distances, will dominate the denominator and the expected LDt will be proportional to the LDC:

$$E\left(LDt\left(\mathbf{b}_k\mathbf{b}_j\right)\right)\propto\left(\beta_j-\beta_k\right)\left(\beta_j-\beta_k\right)^T. \tag{A6}$$

For low noise levels, the first term in Eq. (A4) will dominate the denominator and the expected LDt will be proportional to the square root of the LDC:

$$\begin{aligned}E\left(LDt\left(\mathbf{b}_k\mathbf{b}_j\right)\right)&\propto\left(\beta_j-\beta_k\right)\left(\beta_j-\beta_k\right)^T/\sqrt{\left(\beta_j-\beta_k\right)\left(\beta_j-\beta_k\right)^T}\\&=\sqrt{\left(\beta_j-\beta_k\right)\left(\beta_j-\beta_k\right)^T}.\end{aligned} \tag{A7}$$

In conclusion, this shows that on a continuum from high to low measurement noise, LDt varies in a non-linear fashion between the LDC and its square root, respectively. This non-linear relationship makes it potentially difficult to interpret the ratios between different LDt values.

References

Allefeld, C., Haynes, J.-D., 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage 89, 345–357. http://dx.doi.org/10.1016/j.neuroimage.2013.11.043.

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B., Rätsch, G., 2008. Support vector machines and kernels for computational biology. PLoS Comput. Biol. 4 (10), e1000173. http://dx.doi.org/10.1371/journal.pcbi.1000173.

Bishop, C.M., 2006. Pattern recognition and machine learning. Pattern Recogn. 4, 738. http://dx.doi.org/10.1117/1.2819119.

Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling. Springer Series in Statistics vol. 94 p. 614. http://dx.doi.org/10.1007/0–387-28981-X.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (27), 1–27:27. http://dx.doi.org/10.1145/1961189.1961199.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19 (2), 261–270. http://dx.doi.org/10.1016/S1053-8119(03)00049–1.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis I. Segmentation and surface reconstruction. NeuroImage 9 (2), 179–194.

Diedrichsen, J., Ridgway, G.R., Friston, K.J., Wiestler, T., 2011. Comparing the similarity and spatial structure of neural representations: a pattern-component model. NeuroImage 55, 1665–1678. http://dx.doi.org/10.1016/j.neuroimage.2011.01.044.

Diedrichsen, J., Wiestler, T., Krakauer, J.W., 2013. Two distinct ipsilateral cortical representations for individuated finger movements. Cerebral Cortex (New York, N.Y. : 1991) 23 (6), 1362–1377. http://dx.doi.org/10.1093/cercor/bhs120.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392 (6676), 598–601. http://dx.doi.org/10.1038/33402.

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B.T.T., ... Zilles, K., 2008. Cortical folding patterns and predicting cytoarchitecture. Cereb. Cortex 18, 1973–1980. http://dx.doi.org/10.1093/cercor/bhm225.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 (2), 179–188. http://dx.doi.org/10.1111/j.1469–1809.1936.tb02137.x.

Friston, K.J., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time-series. Hum. Brain Mapp. 1 (2), 153–171. http://dx.doi.org/10.1002/hbm.460010207.

Garrido, L., Vaziri-Pashkam, M., Nakayama, K., Wilmer, J., 2013. The consequences of subtracting the mean pattern in fMRI multivariate correlation analyses. Front. Neurosci. 7, 174. http://dx.doi.org/10.3389/fnins.2013.00174.

Grill-Spector, K., Kourtzi, Z., Kanwisher, N., 2001. The lateral occipital complex and its role in object recognition. Vis. Res. 41 (10–11), 1409–1422 (Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11322983).

Grosenick, L., Greer, S., Knutson, B., 2008. Interpretable classifiers for FMRI improve prediction of purchases. IEEE Trans. Neural Syst. Rehabil. Eng.: A Publication of the IEEE Engineering in Medicine and Biology Society 16 (6), 539–548. http://dx.doi.org/10.1109/TNSRE.2008.926701.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science (New York, N.Y.) 293, 2425–2430. http://dx.doi.org/10.1126/science.1063736.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., ... Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72 (2), 404–416. http://dx.doi.org/10.1016/j.neuron.2011.08.026.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. http://dx.doi.org/10.1146/annurev-neuro-062012-170325.

Henriksson, L., Khaligh-Razavi, S.-M., Kay, K., Kriegeskorte, N., 2015. Visual representations are dominated by intrinsic fluctuations correlated between areas. NeuroImage http://dx.doi.org/10.1016/j.neuroimage.2015.04.026.

Kanwisher, N., McDermott, J., Chun, M.M., Neurosci, J., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. 17 (11), 4302–4311 (Retrieved from http://www.jneurosci.org/content/17/11/4302.full).

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103 (10), 3863–3868. http://dx.doi.org/10.1073/pnas.0600244103.

Kriegeskorte, N., Formisano, E., Sorger, B., Goebel, R., 2007. Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc. Natl. Acad. Sci. U. S. A. 104 (51), 20600–20605. http://dx.doi.org/10.1073/pnas.0705654104.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4. http://dx.doi.org/10.3389/neuro.06.004.2008.

Ku, S., Gretton, A., Macke, J., Logothetis, N.K., 2008. Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. Magn. Reson. Imaging 26 (7), 1007–1014. http://dx.doi.org/10.1016/j.mri.2008.02.016.

Ledoit, O., Wolf, M., 2004. Honey, i shrunk the sample covariance matrix. J. Portf. Manag. http://dx.doi.org/10.3905/jpm.2004.110.

Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605. http://dx.doi.org/10.1007/s10479-011-0841-3.

Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53 (1), 103–118. http://dx.doi.org/10.1016/j.neuroimage.2010.05.051.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10 (4), e1003553. http://dx.doi.org/10.1371/journal.pcbi.1003553.

Op de Beeck, H.P., 2010. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? NeuroImage 49 (3), 1943–1948. http://dx.doi.org/10.1016/j.neuroimage.2009.02.047.

O'Toole, A.J., Jiang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17 (4), 580–590. http://dx.doi.org/10.1162/0898929053467550.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage http://dx.doi.org/10.1016/j.neuroimage.2008.11.007.

Pietrini, P., Furey, M.L., Ricciardi, E., Gobbini, M.I., Wu, W.-H.C., Cohen, L., ... Haxby, J.V., 2004. Beyond sensory images: object-based representation in the human ventral pathway. Proc. Natl. Acad. Sci. U. S. A. 101 (15), 5658–5663. http://dx.doi.org/10.1073/pnas.0400707101.

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage 65, 69–82. http://dx.doi.org/10.1016/j.neuroimage.2012.09.063.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. vol. 8. Springer, p. 188. http://dx.doi.org/10.1109/TNN.1997.641482.

Williams, M.A., Dang, S., Kanwisher, N.G., 2007. Only some spatial patterns of fMRI response are read out in task performance. Nat. Neurosci. 10 (6), 685–686. http://dx.doi.org/10.1038/nn1900.

Williams, M.A., Baker, C.I., Op de Beeck, H.P., Shim, W.M., Dang, S., Triantafyllou, C., Kanwisher, N., 2008. Feedback of visual object information to foveal retinotopic cortex. Nat. Neurosci. 11 (12), 1439–1445. http://dx.doi.org/10.1038/nn.2218.

Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. NeuroImage 5 (3), 179–197 (Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9345548).

Please cite this article as: Walther, A., et al., Reliability of dissimilarity measures for multi-voxel pattern analysis, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.12.012